7
Are We Alone?

7.1 The Need for a Theory

In the introduction to this volume we cautioned against radical theorizing. As a discipline, we probably enjoy inventing new theories too much. Theoretical novelty has become too much of a good thing.

But theories are needed sometimes. As we address the remaining issues of consciousness in animals and robots, direct empirical evidence can only go so far. Some may feel that certain animals are capable of behavior so complex and human-like that it seems absurd to deny that they are conscious. But what if a robot can mimic these exact same behaviors? Is it then conscious too? Not infrequently, the reaction is much more skeptical in that case. We seem to lack a coherent set of objective criteria for making these judgments.

A good way to address this may be to first figure out what *in principle* accounts for consciousness in the human cases. From there, we can see if similar mechanisms exist or not in the animals or robots. This will provide no deductive proof, of course. It assumes that our knowledge of consciousness in humans is correct, *and* that the principles generalize to other creatures. But this is still far better than subjective guesses. To make this inductive generalization, we need a theory of consciousness.

Because neither the global nor local views work well, we need a third option. The theory doesn't have to be brand new; it is more important that it is correct. In the chapter I'll introduce what can be considered a variant of a higher-order theory (Lau and Rosenthal 2011), which I call the perceptual reality monitoring (PRM) theory (Lau 2019). The key ideas can be traced back to John Locke and Immanuel Kant, among others. As I mentioned in Chapter 6, David Rosenthal and Richard Brown are two contemporary champions of variants of this philosophical theory. Rather than proposing something radically new, I will express similar ideas in different terms. But I'll also point out some key differences between our views in Sections 7.4–7.6. The goal is not to contrive originality. Rather, it is to defend that these ideas are generally empirically plausible, and very much compatible with the scientific evidence reviewed in earlier chapters.

In Consciousness We Trust. Hakwan Lau, Oxford University Press. © Hakwan Lau 2022.
DOI: 10.1093/oso/9780198856771.003.0008

7.2 Some Intuitions

On the internet, there are videos of various animals reacting to magic tricks. In one of my favorite demonstrations, a person presented a simple trick to an orangutan. The person placed an object inside a cup, shook it, and then removed the object quickly with a sleight of hand. Afterwards, when the orangutan looked into the cup and couldn't find the object, the animal looked puzzled for a second, and then rolled on the floor with what seemed like bewildered amusement.

Let us assume that we read the emotion of the animal correctly: that the animal was in fact entertained. Can we imagine this animal not having conscious vision? That is, could the animal just be sensing visual information effectively, without having subjective experiences? Perhaps it was something equivalent to a very powerful form of blindsight?

But a blindsight patient may not find magic tricks so entertaining either. Let's say a patient can track the permanence of a moving object, so "guesses" can be made correctly as to whether an object suddenly disappears. Would the patient find this so amazing to watch? Perhaps it is possible the patient may have a gut reaction that something funny is going on when an object suddenly disappears. But to enjoy stage magic involves more than that. When we go to magic shows we hope not just to be nonconsciously "tickled." When we find a magic trick amazing, we enjoy the sense of amazement *as a rational agent*. We find it entertaining in large part because these tricks challenge our grasp of reality. Things appear *crazy*!

So there may be an argument to be made, that there is really no such thing as nonconscious magic tricks. The "unbelievable" nature of magic tricks is a major source of the amusement. But this conflict between our perception and beliefs seems to arise only for conscious perception. In blindsight, the nonconscious perceptual information is in a sense cognitively accessible too, as reflected by the guessing behavior. But the information does not impinge on the patients' belief system the same way as conscious perception does. Blindsight patients don't automatically form firm and rational beliefs when they nonconsciously encounter objects in the world.

Perhaps this lack of direct connection to our beliefs is not specific to blindsight, but true for nonconscious perception in general. If that is so, this may explain why some of us find it so difficult to see how the orangutan could have seen the magic trick only nonconsciously. If the animal lacked conscious vision, the trick just wouldn't have been so genuinely interesting.

7.3 Optimal Bayesians & Phantom Pain

Intuitions are not universal. Unlike me, others may not find it so hard to imagine that a nonconscious animal can somehow enjoy magic tricks too. Or maybe there could be a form of super-blindsight, in which subjective experience is lacking and yet it can directly lead to very *firm* beliefs.

But this firmness of perceptual beliefs deserves further consideration. They say seeing is believing. But as far as conscious seeing is concerned, the result is typically not just some ordinary, casual beliefs, like believing that tomorrow is going to rain. We believe what we consciously see with a certain degree of conviction and immediacy. In most cases, it feels like it is the most reasonable thing to hold on to these beliefs—sometimes even in the face of contradictory evidence. Philosophers sometimes say that perception comes with an "assertoric force." It tells us about what is going on *here and now*, in ways that we can't quite ignore.

This seems to go against the general wisdom for building a rational decision-making system. Engineers and cognitive theorists who take probability theory seriously sometimes identify themselves as Bayesians (after Thomas Bayes's famous theorem on conditional probabilities). These scholars recognize that evidence comes with varying degrees of strength. The optimal way to make decisions is to combine *all* of the different sources of evidence, as weighted by their respective reliability. This is to say, in the face of contradictory evidence, no single source should by default dominate in absolute terms. Let's say I am *very* sure that today is Monday. But if all my friends tell me that today actually is Tuesday, instead, I will probably check the calendar on the internet. If it is confirmed that they are correct, I will revise my belief. I should let the new information override my former conviction and conclude that I was mistaken. The former belief would not retain some mysterious "assertoric force."

Curiously, with conscious perception, this assertoric force seems to never go away. Take the example of phantom pain, which happens in some patients with amputated limbs (Nikolajsen et al. 1997). These patients may feel pain in the limbs that they no longer have. Yet these patients are typically perfectly rational and lucid. They know about the amputation, and the impossibility to have a bodily disturbance in the "location" of the pain per se. And yet they can't *reason* the pain away. They cannot just take into account other evidence and beliefs and let them override and eliminate the "mistaken" pain. The subjective experience remains, and so does the assertoric force. It continues to feel *as if* something is wrong (e.g., being stabbed at) in the location where the pain is felt.

So this assertoric force seems strangely *stubborn*, in the way it has this tendency to inform us about what's happening *here and now*, regardless of what background beliefs we have. With effort we can resist believing in what we currently see, but things would still *seem* to us a certain way, given our conscious percepts. If we are anything like an optimal Bayesian system, the process of evidence accumulation is really not supposed to work this way. As such, perhaps we shouldn't expect nonconscious perception to ever behave like this either. This assertoric force may be a unique and curious feature of consciousness that calls for an explanation.

7.4 Higher-Order Thought or Beliefs?

One way to account for the assertoric force discussed in Section 7.3 is to theorize that conscious perception always involves two sets of representations. We can call the state of early sensory activity the first-order state or first-order representation. This reflects the perceptual content (e.g., what objects are involved and in what spatial location, with specific features like colors, size, and motion direct). These representations are likely within the sensory cortices. We can say they are relatively picture-like, carrying analog content; we will discuss more what "analog" means in Section 9.5.

However, for one to consciously perceive, one may additionally need to have certain higher-order states or representations. That is, the first-order states alone may drive visual behavior and performance. But without the relevant higher-order states, that would only constitute nonconscious perception, as in blindsight. The higher-order states may be reflected by activity in, for example, the prefrontal cortex. The content of these representations may be more conceptual, symbolic, and relatively sentence-like.

What may be the specific content of the higher-order states? Given the discussion in the Section 7.3, one may be tempted to think that these higher-order states could be the corresponding perceptual beliefs. That is, the first-order state contains the picture-like sensory information, for example, about a cat in front of us. For us to consciously see the cat, we need to have the higher-order belief *that* there is the cat in front of us. This belief can then guide our rational decision-making.

But this higher-order "belief" view is too strong. What conscious perception entails is an assertoric force. But this force does not *always* lead to the corresponding belief. For example, some individuals knowingly ingest hallucinogens for recreational purposes. When they hallucinate a cat, they don't necessarily believe that a cat is really out there in front of them. That is, their

background beliefs may ultimately override the assertoric force given by conscious perception. Likewise for the phantom pain example from the last section. Those patients typically don't end up believing that there is bodily disturbance in the location of the felt pain. It just feels *as if* there is such bodily disturbance there. But they know full well that this is impossible in reality because the relevant limb no longer exists.

So what we want is for conscious perception to have the strong *tendency*, but not logical necessity, to lead to the corresponding perceptual belief. Some philosophers have argued for such a "dispositionalist" position (Pitcher 1971). But just stating that there is such a strong disposition doesn't quite make a scientific theory. We need some account of how that disposition comes about.

One influential view is David Rosenthal's higher-order thought theory (2005), which may be an attractive candidate solution here. Continuing with the example of having a first-order representation of a cat, the corresponding higher-order "thought" may have content like: I am having a first-order representation of a cat. Because of this thought, one can account for why one may be disposed to making the corresponding belief; in the absence of contradictory background beliefs, the relevant perceptual belief logically follows.

But if that is the relevant higher-order content, we face another kind of problem. In Section 6.10 we mentioned that the sensory representations may be similar whether it is externally triggered or endogenously generated. These representations are not exactly identical, but seeing a cat and maintaining the image of it in working memory both involve having a first-order representation of a cat. And yet, the mere thought that one is in such a first-order state does not always lead to subjective visual experience. The phenomenology of visual working memory varies across individuals. Some people do not experience imagery during memory delay at all (Zeman, Dewar, and Della Sala 2015). And yet they are no doubt aware of holding the content in mind; they *think* they are in the relevant first-order state—even though there is no corresponding visual experience. Even for those who experience vivid imagery during working-memory delays, the experience is different from normal perception. It lacks that *here and now* quality.

Perhaps this reading of Rosenthal's higher-order thought theory is not so charitable. Maybe the relevant higher-order thought could be more specific: for example, I am seeing a cat *versus* I am holding the image of a cat in working memory. This way we can stipulate that only the former leads to the subjective experience of seeing but the latter doesn't. But the requirements for a philosophical theory may be different from what we need for

a scientific theory. Given our purpose here, we should have a mechanistic explanation for how these different higher-order states come about. Just saying that these may be the possible contents in abstract terms isn't very satisfying. Ideally, we should be able to describe them in enough detail so that we can in principle *build* a system capable of generating these higher-order states.

7.5 Inner Sense

In Section 6.10, we introduced generative adversarial networks (GANs). To facilitate the development of a system capable of predictive coding, a "discriminator" may be employed. The job of this discriminator is exactly to distinguish between the different kinds of first-order states described in the last section: endogenously generated versus externally triggered. So we can think of the discriminator as a higher-order mechanism for consciousness.

One interesting feature of the discriminator is that it is basically just a simple *perceptual* categorization network. In this sense the higher-order process itself isn't exactly thought—or sentence-like (although it may output to further downstream processes that are more so). So, this may correspond to another variant of higher-order theory known as "inner-sense" theory, or higher-order *perception* theories. These theories have been criticized on philosophical grounds (Carruthers 2007), but I'm not sure the arguments are decisive.

In particular, one challenge raised by critics of the "inner-sense" theory is empirical (Sauret and Lycan 2014). They argued that no such inner-sense "organ" has been found. But today we know that there may be good computational reasons why the brain would have employed a GANs-like architecture. The neurophysiological evidence suggests that such a mechanism may be in the prefrontal cortex (Section 6.10). And then computationally, it seems like the same mechanism could be repurposed for metacognition. This in turn neatly explains why disrupting activity in the prefrontal cortex, where such discriminator function likely locates, can impair metacognition too (Chapter 3).

I mentioned the above "repurposing" finding as a result of a computational modeling exercise (Section 6.10). But conceptually we can also understand why the discriminator may be useful for metacognition. In general, if one becomes very good at distinguishing between two subtypes of X, one tends also to be better at detecting X from non-X. For example, if we are really good at telling red wines from white wines by taste, chances are we can detect just in a

sip if some wine has been added to a glass of water too. So likewise, since the discriminator is capable of distinguishing between internally generated and externally triggered first-order states, it makes sense that it can distinguish the presence of a perceptual signal from sheer noise too.

This metacognitive function—of distinguishing between a meaningful perceptual signal and noise—is important because spontaneous neuronal activity is ubiquitous. It accounts for much of the metabolic budget of the brain (Raichle 2006; Schölvinck, Howarth, and Attwell 2008). As I'm writing, my cat-representing neurons, like most other sensory neurons, fire now and then. But I do not hallucinate seeing cats (!). Somewhere there must be a mechanism in the brain for deciding that such spontaneous activity is just "noise," rather than caused by a cat in front of me.

Let's assume that an agent is capable of reasoning with beliefs and goals. If the mechanisms for this kind of general symbolic-level cognition receive input from both the first-order states and the discriminator, one can see how conscious perception can attain its assertoric force. Essentially, conscious perception happens when the first-order state represents the cat, *and* the higher-order (discriminator) state indicates that the first-order state is a true reflection of the world right now. Together, these two representations constitute something akin to the premises of a syllogistic inference. In the absence of conflicting background beliefs, it is rational to form the belief that there is a cat in front of us right now. Such a belief is in a sense "justified" by the premises; it logically follows (Figure 7.1).

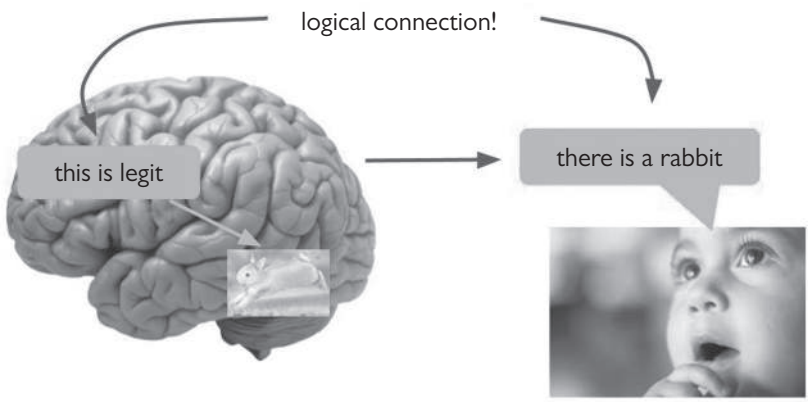

Figure 7.1 Perceptual reality monitoring via first- and higher-order representations

7.6 Index, Gating, & the Richness of Experience

The last point sets the theory apart from most other versions of higher-order theories because here perceptual beliefs are derived from both the first-order (sensory) and higher-order (discriminator) states. In higher-order thought theory, the content of consciousness is ultimately determined by the higher-order state (Rosenthal 2005; Lau and Brown 2019). The first-order state is the normal cause or input to the higher-order state, but it isn't constitutively part of the subjective experience. Once the higher-order state is formed, in principle, both the subjective experience and the relevant perceptual beliefs can occur with or without the first-order state.

Against the higher-order thought view, there may be some concerns regarding whether the thought-like higher-order representation can capture the richness of perceptual experience. As reviewed in Chapter 4, the issue is controversial. But a standard higher-order *thought* theorist is committed to a relatively "sparse" view, on which perceptual experience is no richer than what can be captured by conceptual, thought-like representations. The view argued for here makes no such commitment; the rich content of the first-order state also contributes. The higher-order state does not duplicate the first-order content, but merely serves as a gating mechanism to direct the first-order information to the relevant downstream processes.

The reader may wonder: how does the higher-order (discriminator) state *refer* to the corresponding first-order state, without duplicating its content? In current artificial GANs, we tend to deal with just one first-order state at a time. But in the actual human brain, there may be multiple concurrent perceptual states in different sensory modalities. At a given time, some of these first-order states may lead to subjective experience while others may not. As such, there are likely multiple discriminator outputs, and one needs to keep track of which refers to which first-order states. Of relevance is that *indexing* or variable binding mechanisms have been proposed for prefrontal functions (Kriete et al. 2013). Essentially, the prefrontal cortex must have some ways of referring to specific first-order activities via some form of "addressing" system.

For a simplistic analogy, we can think of this as a phone numbers system, where each individual referent is given a unique identifier. So, as in modern computational systems, a higher-order mechanism can refer to first-order representations by these addresses, without duplicating or redescribing the full content. In Chapter 9 we will revisit how such mechanisms may actually work in the mammalian sensory cortices.

For these indexes or addresses to work, they need to be interpretable by some downstream system. Such a system must be able to access both the

higher-order and first-order content. On the view advocated here, the idea is that these contents are subsequently read out by the mechanisms for general symbolic-level cognition and logical reasoning. In this sense, consciousness is the *gating mechanism* by which perception impacts cognition; it selects what perceptual information should directly influence our rational thinking.

Note that this is not to identify consciousness with access consciousness (as discussed in Section 1.6). *Consciousness* here refers to subjective experience, as we do throughout most of this book. The point is that subjective experiences are *causally* connected with access consciousness in the ways described above. Subjective experiences are characterized by their *availability* for *potential* conscious access. But I'm not suggesting that the two are one and the same. When the discriminator decides that a first-order representation correctly represents the world right now, global broadcast and access are *likely* to happen. But these consequences are not constitutively part of the subjective experience, according to this view.

7.7 Phenomenology of Imagery

We can call the view introduced in the last few sections the PRM theory. It is so-called because in the memory literature, a similar process of reality monitoring has been proposed (Johnson and Raye 1981; Johnson 1988). For example, young children and older adults alike sometimes confuse their own past imagination with events that actually took place. This could lead to dire consequences, if they were in fact mistaken and to bear witness in court cases, for example. Fortunately, this doesn't happen more often, because monitoring mechanisms exist in the brain to determine the source of a memory. This allows us to tell apart reality from fantasy, in a generally reliable fashion. By identifying such reality monitoring mechanisms in the prefrontal cortex (Simons, Garrison, and Johnson 2017), we can also in principle assess how trustworthy one's witness statements may be, based on neurological data.

Likewise, in the case of perceptual signals, there is a similar need to distinguish between reality and our own mental imagery. But does PRM imply that endogenously generated perceptual signals are always nonconscious? The short answer is *no*. But it is somewhat complicated.

What is clear is that normal functioning subjects do not generally confuse normal perception with imagery; the phenomenology is typically distinct in the two cases. There has been some scant evidence that such confusion is common, but the results, at least in their original forms, are not robustly replicable (Segal and Nathan 1964; Segal and Gordon 1969). Modern empirical

studies have found relatively subtle interference between perception and endogenously maintained perceptual information (Kang et al. 2011; Salahub and Emrich 2016; Teng and Kravitz 2019; Dijkstra and Fleming 2021; Dijkstra et al. 2021; Dijkstra, Kok, and Fleming 2021).

But what is the phenomenology for endogenously generated perceptual signals? Is it distinct from normal perception because it is "weaker" or absent? There are considerable individual differences regarding the phenomenology of mental imagery. In aphantasia, the concerned individuals do not experience visual imagery in vivid forms at all (Zeman, Dewar, and Della Sala 2015). One possibility is that the relevant first-order sensory states are either absent or only partially instantiated when these subjects engage in visual thinking. But how come one can still perform visual functions without these first-order activities?

According to PRM, there is another possibility why aphantasia or weak imagery experience may happen. We have argued that the neural mechanisms for the discriminator may also contribute to metacognition. So in total, this discriminator-like mechanism has *three* different output conditions. That is, a first-order state is one of the following: i) externally triggered, ii) internally generated, or iii) just noise. The first condition should lead to normal subjective perceptual experience, and that the third condition should entail the lack of subjective experience. When the discriminator decides that a certain first-order state is internally generated (second condition), a distinct output is needed. Whether this output is more similar to the first or the third condition may vary across people; when we say two outputs are similar, we mean that they are more easily confused to be the same by downstream readout. To the extent that it is more similar to the third condition (noise) rather than the first (externally triggered), PRM predicts that subjective experience may be absent or relatively feeble too—even if the first-order activity is actually robust.

Incidentally, although neurophysiological and anatomical correlates of imagery vividness have been found in sensory areas, similar findings have also been reported for prefrontal and parietal areas (Dijkstra, Bosch, and van Gerven 2019). The relative paucity of evidence for the higher-order areas may again be due to methodological considerations already discussed in Chapters 2 and 3. As such, one may hypothesize that imagery vividness can be causally manipulated if we tamper with *either* the first-order or the higher-order states (appropriately). I am not aware of this being formally tested yet: to induce confusion between imagery and perception with prefrontal or parietal stimulations. PRM hereby makes this empirical prediction.

7.8 Implicit Versus Explicit Reality Monitoring: The Case of Dreams

The last point highlights why PRM is a balanced synthesis between local and global theories. Local theorists emphasize the causal relevance of early sensory mechanisms. They certainly have a point, but that picture is empirically incomplete. In reaction, global theorists sometimes come across as putting too much emphasis on the prefrontal and parietal cortices alone. But both the higher-order and first-order states are important. Various "disorders" of consciousness can be caused by abnormalities in either (Zmigrod et al. 2016). A correct theory must be able to account for both.

The common phenomenon of dreaming is typically not considered a "disorder." But in dreams we are typically "mistaken" in a sense: we treat our internally generated sensory activities as reflecting the present state of the world. In other words, dreams are a form of hallucinations. As in other kinds of hallucinations, there are no doubt first-order correlates (i.e., activities in the sensory regions of the brain; Horikawa et al. 2013). But as we have pointed out, working memory and mental imagery also involve similar early sensory activities. And yet in dreams the sensations seem far more vivid. One may argue that during working memory and mental imagery the sensory activities are perhaps not as strong and detailed. But in dreams these activities are also generated endogenously. In all these cases, concurrent external input is lacking. What sets the corresponding subjective experiences apart?

Although not all dreams happen during REM (rapid eye movements) sleep, we dream more often during REM than non-REM sleep. Incidentally, during REM sleep sensory cortices tend to be active, and yet prefrontal areas like the dorsolateral prefrontal cortex often show reduced levels of activity (Muzur, Pace-Schott, and Hobson 2002). This has been taken as a challenge to traditional versions of the higher-order view; perhaps less activity in the prefrontal cortex means fewer higher-order thoughts, and those one should not be having vivid experiences as we do in dreams. But as we explained in Chapter 3, neurons in the prefrontal cortex do not fire to signal simply the presence of a stimulus. Instead, they form complex high-dimensional neuronal population codes. As such, we need to be careful in interpreting the lack of salient prefrontal activity as observed by crude neuroimaging measures during REM sleep. One plausible interpretation is that during REM sleep the prefrontal areas may be *failing* their usual role in PRM, leading us to mistake endogenously generated sensory activities as caused directly by the external world. The low activity exactly reflects the disengagement of the relevant process.

Incidentally, we are not always mistaken about the nature of dreams. Occasionally, people have lucid dreams, in which they know full well that they are dreaming. They know that their subjective experiences are detached from reality. Lucid dreaming has been associated with heightened activity in the prefrontal cortex (Dresler et al. 2012; Stumbrys, Erlacher, and Schredl 2013). Perhaps this is why we recognize the illusory nature of dream percepts during lucid dreams.

However, an important distinction needs to be made. In lucid dreams we continue to have vivid subjective experiences. The correct and explicit monitoring of reality happens at a higher cognitive level. This is similar to the case of known hallucinations described in Section 7.4. Even if we don't ultimately form the mistaken beliefs, the subjective experiences in dreams and known hallucinations are still somewhat misleading, as they carry this undeserved "assertoric force." So for these instances, we can say that reality monitoring fails at this implicit, subpersonal, and automatic level, even though this is corrected downstream at the higher cognitive, explicit level.

For PRM and subjective experience, it is this *implicit* kind of reality monitoring that really matters. The idea is that the prefrontal cortex may be important for *both* kinds of reality monitoring, implicit (i.e., subpersonal) as well as explicit (i.e., higher cognitive). If one fails, the other may not; the relevant circuits need not be exactly identical, even though they may partially overlap, and may both reside within the prefrontal cortex. The partial overlap between the explicit and implicit functions may explain why lucid dreaming is possible but relatively rare; it requires the explicit reality monitoring to function but implicit reality monitoring to fail.

There is some evidence in support for the idea that the prefrontal cortex may be important for both kinds of metacognition in dreams, explicit and implicit, although the mechanisms may be ultimately different. Applying transcranial electrical stimulation through the scalp, Voss et al (2014) have reported that targeting the prefrontal cortex can increase incidents of lucid dreaming (explicit metacognition). However, such stimulation also seems to have effects on the reported frequency of dreams, as well as the reported degree of "realism" of the sensory details in dream content (see the supplementary tables in Voss et al. 2014). Subjects seem to more frequently mistake endogenously generated "noise" as reflecting the outside world (*failure* of implicit metacognition).

Based on electroencephalogram (EEG) data, others have suggested that some regions in the medial parietal areas may be important for subjective experiences to arise in dreams too (Siclari et al. 2017). This is not in contradiction with higher-order theories; although the prefrontal cortex is often emphasized, both global and higher-order theorists actually also recognize

the importance of other areas in the association cortices, including parietal areas (see Section 3.1). But one concern is that some of these areas, such as the precuneus, are also linked to memory metacognition (as reviewed in Chapter 3) and the vividness of memory recall (Richter et al. 2016). So one needs to make sure the activity truly reflects subjective experience *per se*, rather than just the memory and subsequent reportability of these experiences.

One should also note that in the EEG study mentioned above, prefrontal activities were also found for subsequently reported dreams (Siclari et al. 2017). The researcher wrote off such activity as relatively weak and inconsistent. But as in the findings on REM sleep mentioned above, such weak signals may actually reflect the failure of subpersonal PRM, which can be the causal explanation for subjective experiences in dreams. Also, with a relatively crude method like EEG we should not overinterpret null or weak results. Imaging methods do not give the same sensitivity to different signals in different regions, as we have already discussed in Chapter 3. Fazekas and Nemeth (2018) provide a useful review suggesting that the emphasis on medial parietal areas at the expense of prefrontal involvement may be empirically unsound.

7.9 Other Higher-Order Failures

Why do we think that both explicit and implicit PRM depend on the same brain regions? I take it that for the explicit, higher-cognitive variant, it is not controversial that the prefrontal cortex is important. Failure of reality monitoring at this level amounts to general delusions. The prefrontal cortex is known to be important for normal cognitive functioning and reasoning.

But just because explicit metacognition and explicit PRM depend on the prefrontal cortex does not mean that the relevant implicit processes must be carried out in earlier sensory regions. A single cortical area often subserves multiple functions. Given the similarity in the overall computational goal, it would make sense that the implicit and explicit circuits are in close proximity or may partially overlap. Assuming explicit metacognition and reality monitoring are evolutionarily "newer" functions, they may be built upon the relevant implicit mechanisms, "recycling" some of the same neural resources and similar circuits. But it does not mean that the explicit and implicit functions are the same.

Phil Corlett has made the argument that higher-cognitive delusions and sensory hallucinations may not be driven by totally independent factors, as they sometimes suggested (2019). Many current theories of psychosis identify dopaminergic functions in the prefrontal cortex as key mechanistic

components (Braver, Barch, and Cohen 1999). Again, this is not to say that when patients with schizophrenia hallucinate, early sensory activities are irrelevant. The point is that perhaps the higher-order mechanisms in the prefrontal cortex are important too, even in the absence of delusions. The prefrontal cortex may be the common factor in *both* sensory-level hallucinations as well as higher-cognitive delusions. We will address further in Chapter 8 how malfunctions at the two levels may interact.

An often overlooked fact is that some patients with Parkinson's disease or Lewy body dementia also hallucinate visually (Onofrj et al. 2013). Unlike patients with schizophrenia, they are relatively lucid and cognitively intact. Again, Parkinson's disease is characterized by impairments of dopaminergic functions in the frontal lobes, including the subcortical basal ganglia (Narayanan, Rodnitzky, and Uc 2013).

Finally, as we already discussed in Chapter 2 Section 2.11, even for a disorder of subjective experience caused by lesion to the primary visual cortex—blindsight—there are clear physiological correlates in the prefrontal cortex. Theoretically it has been suggested that higher-order malfunctioning is the ultimate culprit (Ko and Lau 2012).

7.10 Inflation Revisited

Section 7.9 may help us understand the possible role of prefrontal mechanisms in inflation. Recall from Chapter 4 that *inflation* refers to the occurrence of subjective experience under the relative lack of representational details. This may happen in peripheral vision, for instance.

Suppose we have the first-order sensory state of a cat that is rather impoverished. As a pictorial representation it lacks details. The content barely reflects a real cat. There is no way to tell if it is a Persian cat or a Bengal, or neither. So, an unbiased optimal system may well consider this to be noise, rather than a truthful representation of the current state of the world. The representation just isn't reliable enough to be taken seriously.

But according to PRM, the relevant higher-order mechanism is distinct from the first-order state. If the higher-order mechanism *somehow* makes the judgment that the impoverished first-order state correctly reflects the current world, according to the theory there will be a corresponding subjective experience. Instead of being filtered out from further processing, the weak first-order representation would be available for higher-cognitive access. There will be an assertoric force that comes with the conscious experience, to the effect that "something like a cat" (or whatever represented by the first-order

state) is in front of the subject. This may explain why in peripheral or unattended vision, there is a robust liberal detection bias, even when one is not so able to discriminate or identify the relevant target well (Section 4.9). The perceptual content ultimately lacks the detail, so upon reflection it may not be "rich" per se. But if it is being deemed by the higher-order mechanism as rich enough, there is a sense in which the perceptual experience is subjectively rich or strong; the subject is likely to form the corresponding belief with certainty.

Why would the higher-order mechanism make such a biased judgment, given that the first-order state isn't quite detailed and robust enough to be reliably distinguished from noise? One reason is that the higher-order mechanism may err, as we have discussed in the Sections 7.8 and 7.9. But such an "error" may also be a useful heuristic. For example, in peripheral or unattended vision, the first-order state is *expected* to lack certain details. But that is not because *the world* lacks such details. Rather, our brains should "know" that such details are just a saccade away; we only have to look. So despite the lack of richness in the first-order content for the unattended periphery at one moment, the higher-order mechanism may reasonably give such content some "advanced" credit.

This account of inflation may apply also to the phenomenology of dreams. Although dreams feel vivid, upon reflection it is often unclear if all the details are really there. Perhaps it only *feels as if* the rich details are there—just as in peripheral or unattended vision.

7.11 Agency & Emotions

Although much of the evidence and analysis come from vision, PRM is meant to apply to all sensory modalities. But what about other experiences such as volition and emotions?

One could envision something like this: like perceptual representations, representations for action in the motor cortex are also activated in different ways (Gallese and Goldman 1998; Oosterhof, Tipper, and Downing 2012; Taube et al. 2015; Zabicki et al. 2017). When one imagines acting a certain way, there are similar neural activities as if one is performing the same action. When one observes another person making the action, similar activities also arise. And of course, spontaneous neural activity is ubiquitous. So there is a need for volitional reality monitoring too.

Likewise for affective reality monitoring: experiencing an emotion seems to activate some similar neural activities as when one is merely imagining it, or thinking about another person experiencing it (Sato et al. 2004; Singer et al.

2004). And of course neurons in affect-related brain regions such as the amygdala and insular also show spontaneous neural activity. So perhaps a similar discriminator or reality monitoring mechanisms may be at work in the prefrontal cortex.

But I suspect that there is more to it for the sense of agency and emotional experiences. These conscious experiences are in a sense more complex, as they involve a more explicit notion of the self (LeDoux and Lau 2020). Higher-cognitive and memory mechanisms may contribute to these experiences more than they do for simple perceptual experiences. We will address these possibilities further in Chapter 8.

7.12 Other Minds

With the above caveats in mind, we can finally address the question we started with in this chapter: besides us humans, what else is conscious? Specifically, let us set aside the potentially more complicated question regarding the sense of volitional control and emotions. What creatures are capable of having the simplest conscious perceptual experiences?

I started off with stage magic as an intuitive example. But of course that would not constitute a universally practical test for consciousness. One may need to be capable of having subjective perceptual experiences in order to appreciate stage magic, but other cognitive abilities may also be required.

To determine consciousness we need to get at the precise mechanisms. The relevant creature should first be capable of predictive coding, in a specific way. In particular, when the system generates sensory activity in a top-down manner, it should make use of the same machinery for bottom-up perception. This creates a need for a mechanism akin to a "discriminator" in GANs. This discriminator also has to be capable of metacognition (i.e., to distinguish meaningful sensory representations from noise). Finally, there needs to be a general reasoning and belief-formation system to which the discriminator signals.

This kind of sensory predictive coding mechanism seems present in many mammals. But when it comes to the discriminator function, it is far less clear. There is some evidence that rats are capable of some degree of metacognition, and the mechanisms may also depend on the prefrontal cortex (Stolyarova et al. 2019). But are these the same mechanisms for PRM? It is known that the rodent and primate prefrontal cortices are markedly different, in terms of basic anatomy (Schaeffer et al. 2020).

Related to this was the finding that some neurons can distinguish between perceptual and working-memory content (Mendoza-Halliday and Martinez-Trujillo 2017). These neurons are found in a prefrontal region known to be important for metacognition (i.e., the dorsolateral prefrontal cortex) in monkeys. So far, we lack direct evidence in smaller animals.

And are rats capable of reasoning with their beliefs and desires? Do they really have a general cognitive system for rational thoughts, like we have?

Despite these unresolved issues, I'm inclined to think that most mammals capable of sensory predictive processing and metacognition are probably having some simple conscious experiences. That leaves many smaller animals out. I do not feel certain about this but at least we can spell out the sources of this uncertainty, as I did above.

How about young children and babies? It has been shown that preverbal infants are capable of some degree of metacognition (Goupil and Kouider 2016, 2019; Goupil, Romand-Monnier, and Kouider 2016). As in the case of smaller animals, it is not entirely clear if they make use of the exact same mechanisms for PRM. But because of their developmental trajectory the case may be stronger here. Also, although young children aren't capable of sophisticated reasoning with counter-factual beliefs, they are probably capable of some rational thinking based on beliefs and desires. So according to PRM they are likely conscious too.

That is to say, although some details may be currently not fully proven, these are empirically addressable issues. Assuming PRM is right, we can look for whether human infants and other animals have the essential mechanisms. This is a method of induction. We are assuming that PRM is correct not only in the subjects we have tested. We are hoping it generalizes. But this may be the best that we can do.

The more interesting case may be robots. There we do not have the same uncertainty driven by the lack of empirical data. Many current neural network models are capable of predictive processing. But typically, the generative model projects its top-down outputs to a set of nodes distinct from those used for bottom-up perception (Pu et al. 2016). So they do not have the same pressure to avoid the confusion. That said, some current models are already more brain-like, with the same sensory circuits being used in both top-down and bottom-up processing (Rasmus et al. 2015). Also, of course, the very notion of the discriminator comes from neural network models.

The more challenging part may concern belief formation. Artificial General Intelligence, that is a computational system capable of human-like rational cognition, is a challenging goal (Goertzel 2014). While some current systems

are capable of problem-solving and decision-making, it is not clear to what extent they resemble our cognitive architecture. In particular, it is unclear if such systems are truly capable of producing "thoughts" and "beliefs" in the general sense.

But at least in the current version of PRM, the commitment of the theory should be clear: to the extent a robot has such general reasoning capacities like we do, and if a discriminator signals to such mechanisms that a certain sensory representation is a truthful reflection of the current world, the robot will form a "conscious experience" of the relevant sensory content.

7.13 The Hard Enough Problem

The last point may sound so wild that some may see it as an exposition of the problem of the theory. A robot is just a machine. How can something as special as subjective experience come out of sheer computations?

I agree with this sentiment. Together with other authors I have previously speculated on the issue of machine consciousness (Dehaene, Lau, and Kouider 2017). But I have come to think that our proposal was unsatisfactory. Some key elements seem to be still missing. Perhaps in this preliminary version of PRM, it is not clear how the sense of self comes about, which may be important for at least some forms of conscious experiences (LeDoux and Lau 2020). In Chapter 8 we will try to address that.

But there is also a widespread intuition that the relevant substrate may also matter; this is what ultimately motivates biopsychism (as introduced in the Chapter 6). Machines made of electronics rather than wet, living brains, just seem incapable of *feeling what it is like* to be in certain subjective experiences. Maybe this is just an intuition. But I promise I will try to give my best shot at addressing this issue in the final chapter too.

For now, however, it is important for us to realize that this is a problem for all theories of consciousness. Once a mechanism of consciousness is spelled out, we can try to imagine building a simple creature just barely having the mechanism and proceed to consider if such a simple creature is plausibly conscious. Instead of fixating on the implausibility in absolute terms, we would do well to see how other theories fare and compare accordingly. We can call this the Hard Enough Problem. By "hard enough," I don't mean it is "pretty hard"; I mean that it is exactly hard *enough* for our purpose *of arbitrating between different theories*. The least implausible theory may be the best we can ever have.

Let us consider the robot described in Section 7.12 in more detail. It is true that it is not flesh and blood. But suppose it has some sensors for detecting

bodily damage. When the discriminator tells its general reasoning system that the relevant sensory activity is correctly reflecting the world right now, it will form the belief that something in a certain part of the body is damaged. Suppose it is a false alarm; upon checking, that part of the body actually looks fine and functions well. But because of the way the discriminator is connected to the system for general cognition, it will continue to have this unshakable assertoric force, that something is wrong in that specific part of the body. It can't reason that signal away. That signal will continue to impinge on its rational thinking, *as if* that part of the body is damaged.

How different is this from pain as we know it?

But let us also consider similar cases for other theories of consciousness. For the global view, all it takes is the global broadcast of information, through some central information exchange system. Many current computer network systems already have such mechanisms. Are they conscious then?

And for the local theorist, if the right kind of local sensory activity is really all it takes, what if we isolate such activity so it does not make any downstream impact to other brain areas? What if we keep the relevant neurons in vitro, on a petri-dish, and stimulate them to mimic normal activity? Would they be conscious?

So, all current theories seem to make some rather improbable predictions. Of course, if we know *for sure* that a theory is correct, we should accept whatever improbable consequences it entails. But I hope the reader should have been convinced by now, that the science of consciousness is just no such simple matter. It would take some profound lack of critical thinking for one to accept that *any* current theory can be considered absolutely proven at the moment—including PRM, of course. So the Hard Enough Problem matters. And *perhaps* PRM offers one of the least implausible solutions for now.

If this doesn't feel quite plausible enough just yet, maybe something is in fact missing still. I hope that Chapter 8, and especially also Chapter 9, may convince you a bit more.

7.14 Chapter Summary

Consciousness in animals and robots is not an easy topic. Unfortunately, on this issue, scientists have often made premature and grandiose claims. These claims are rarely based on evidence and logic. Rather, philosophically unexamined intuitions masquerade as established scientific viewpoints. Here I try to make the case that this really should be a two-way process. First, we should see what the most empirically plausible theory says. Then in turn

we should also evaluate the theory based on what the theory says about the matter. If it is just too outlandish, perhaps it would be grounds for rejecting the theory.

The theory I advocate, PRM, suggests that conscious experiences arise out of self-recognized perception of some sort. If a specific inner-sense mechanism "perceives" a sensory representation to be correctly reflecting the world at present, we become conscious of the sensory content. Because this mechanism directly impacts our rational thinking, consciousness can be understood as the *interface* between perception and cognition. Subjective experiences are the things that we are naturally inclined to believe—and *trust*.

PRM faces challenges from the Hard Enough Problem. Many may find it counter-intuitive that a robot can ever be conscious in the sense of having subjective, qualitative experiences. But the logic of the Hard Enough Problem is that it is a relative matter. It depends on what other theories say, which is often far more improbable.

So perhaps this is good enough. But some readers may feel that the theory still does not get at the qualitative aspect of *what it is like* to have a certain conscious experience. We will address this problem in Chapter 9, in order to finally give a better answer to the problem of machine consciousness. But before that, we need to first place the problem in a broader context, to see what really is at stake. In doing so, we will also expand the theory a bit in order to account for emotion and the subjective sense of agency, which are of course no less important than simple perceptual experiences.

References

Braver TS, Barch DM, Cohen JD. Cognition and control in schizophrenia: A computational model of dopamine and prefrontal function. *Biol Psychiatry* 1999;**46**:312–328.

Carruthers P. Higher-order theories of consciousness. *The Blackwell Companion to Consciousness*. Wiley-Blackwell Publishers, 2007;**10**:9780470751466.

Corlett PR. Factor one, familiarity and frontal cortex: A challenge to the two-factor theory of delusions. *Cogn Neuropsychiatry* 2019;**24**:165–177.

Dehaene S, Lau H, Kouider S. What is consciousness, and could machines have it? *Science* 2017;**358**:486–492.

Dijkstra N, Bosch SE, van Gerven MAJ. Shared neural mechanisms of visual perception and imagery. *Trends Cogn Sci* 2019;**23**:423–434.

Dijkstra N, Fleming SM. Fundamental constraints on distinguishing reality from imagination. 2021. https://doi.org/10.31234/osf.io/bw872.

Dijkstra N, Kok P, Fleming SM. Perceptual reality monitoring: Neural mechanisms dissociating imagination from reality. 2021. https://doi.org/10.31234/osf.io/zngeq.

Dijkstra N, Mazor M, Kok P et al. Mistaking imagination for reality: Congruent mental imagery leads to more liberal perceptual detection. *Cognition* 2021;**212**:104719.

Dresler M, Wehrle R, Spoormaker VI et al. Neural correlates of dream lucidity obtained from contrasting lucid versus non-lucid REM sleep: A combined EEG/fMRI case study. *Sleep* 2012;**35**:1017–1020.

Fazekas P, Nemeth G. Dream experiences and the neural correlates of perceptual consciousness and cognitive access. *Philos Trans R Soc Lond B Biol Sci* 2018;**373**. https://doi.org/10.1098/rstb.2017.0356.

Gallese V, Goldman A. Mirror neurons and the simulation theory of mind-reading. *Trends Cogn Sci* 1998;**2**:493–501.

Goertzel B. Artificial general intelligence: Concept, state of the art, and future prospects. *J Artif Gen Intell* 2014;**5**:1–48.

Goupil L, Kouider S. Behavioral and neural indices of metacognitive sensitivity in preverbal infants. *Curr Biol* 2016;**26**:3038–3045.

Goupil L, Kouider S. Developing a reflective mind: From core metacognition to explicit self-reflection. *Curr Dir Psychol Sci* 2019;**28**:403–408.

Goupil L, Romand-Monnier M, Kouider S. Infants ask for help when they know they don't know. *Proc Natl Acad Sci* 2016;**113**:3492–3496.

Horikawa T, Tamaki M, Miyawaki Y et al. Neural decoding of visual imagery during sleep. *Science* 2013;**340**:639–642.

Johnson MK. Reality monitoring: An experimental phenomenological approach. *J Exp Psychol Gen* 1988;**117**:390–394.

Johnson MK, Raye CL. Reality monitoring. *Psychol Rev* 1981;**88**:67–85.

Kang M-S, Hong SW, Blake R et al. Visual working memory contaminates perception. *Psychon Bull Rev* 2011;**18**:860–869.

Ko Y, Lau H. A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. *Philos Trans R Soc Lond B Biol Sci* 2012;**367**:1401–1411.

Kriete T, Noelle DC, Cohen JD et al. Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proc Natl Acad Sci U S A* 2013;**110**:16390–16395.

Lau H. Consciousness, metacognition, & perceptual reality monitoring. 2019. https://doi.org/10.31234/osf.io/ckbyf.

Lau H, Brown R. The emperor's new phenomenology? The empirical case for conscious experiences without first-order representations. In: A Pautz, D Stoljar (eds), *Blockheads! Essays on Ned Block's philosophy of mind and consciousness*. The MIT Press; 2019; 171–197.

Lau H, Rosenthal D. Empirical support for higher-order theories of conscious awareness. *Trends Cogn Sci* 2011;**15**:365–373.

LeDoux JE, Lau H. Seeing consciousness through the lens of memory. *Curr Biol* 2020;**30**:R1018–R1022.

Mendoza-Halliday D, Martinez-Trujillo JC. Neuronal population coding of perceived and memorized visual features in the lateral prefrontal cortex. *Nat Commun* 2017;**8**:15471.

Muzur A, Pace-Schott EF, Hobson JA. The prefrontal cortex in sleep. *Trends Cogn Sci* 2002;**6**:475–481.

Narayanan NS, Rodnitzky RL, Uc EY. Prefrontal dopamine signaling and cognitive symptoms of Parkinson's disease. *Rev Neurosci* 2013;**24**. https://doi.org/10.1515/revneuro-2013-0004.

Nikolajsen L, Ilkjaer S, Krøner K et al. The influence of preamputation pain on postamputation stump and phantom pain. *Pain* 1997;**72**:393–405.

Onofrj M, Taylor JP, Monaco D et al. Visual hallucinations in PD and Lewy body dementias: Old and new hypotheses. *Behav Neurol* 2013;**27**:479–493.

Oosterhof NN, Tipper SP, Downing PE. Visuo-motor imagery of specific manual actions: A multi-variate pattern analysis fMRI study. *Neuroimage* 2012;**63**:262–271.

Pitcher G. *Theory of Perception*. Princeton University Press, 1971.

Pu Y, Gan Z, Henao R et al. Variational autoencoder for deep learning of images, labels and captions. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). Curran Associates Inc., Red Hook, NY, USA, 2360–2368.

Raichle ME. Neuroscience. The brain's dark energy. *Science* 2006;**314**:1249–1250.

Rasmus A, Valpola H, Honkala M et al. Semi-supervised learning with Ladder networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'15). MIT Press, Cambridge, MA, USA, 3546–3554.

Richter FR, Cooper RA, Bays PM et al. Distinct neural mechanisms underlie the success, precision, and vividness of episodic memory. *Elife* 2016;**5**. https://doi.org/10.7554/eLife.18260.

Rosenthal D. *Consciousness and Mind*. Oxford University Press, 2005.

Salahub CM, Emrich SM. Tuning perception: Visual working memory biases the quality of visual awareness. *Psychon Bull Rev* 2016;**23**:1854–1859.

Sato W, Kochiyama T, Yoshikawa S et al. Enhanced neural activity in response to dynamic facial expressions of emotion: an fMRI study. *Brain Res Cogn Brain Res* 2004;**20**:81–91.

Sauret W, Lycan WG. Attention and internal monitoring: A farewell to HOP. *Analysis* 2014;**74**:363–370.

Schaeffer DJ, Hori Y, Gilbert KM et al. Divergence of rodent and primate medial frontal cortex functional connectivity. *Proc Natl Acad Sci U S A* 2020;**117**:21681–21689.

Schölvinck ML, Howarth C, Attwell D. The cortical energy needed for conscious perception. *Neuroimage* 2008;**40**:1460–1468.

Segal SJ, Gordon PE. The perky effect revisited: blocking of visual signals by imagery. *Percept Mot Skills* 1969;**28**:791–797.

Segal SJ, Nathan S. The perky effect: Incorporation of an external stimulus into an imagery experience under placebo and control conditions. *Percept Mot Skills* 1964;**18**:385–395.

Siclari F, Baird B, Perogamvros L et al. The neural correlates of dreaming. *Nat Neurosci* 2017;**20**:872–878.

Simons JS, Garrison JR, Johnson MK. Brain mechanisms of reality monitoring. *Trends Cogn Sci* 2017;**21**:462–473.

Singer T, Seymour B, O'Doherty J et al. Empathy for pain involves the affective but not sensory components of pain. *Science* 2004;**303**:1157–1162.

Stolyarova A, Rakhshan M, Hart EE et al. Contributions of anterior cingulate cortex and basolateral amygdala to decision confidence and learning under uncertainty. *Nat Commun* 2019;**10**:4704.

Stumbrys T, Erlacher D, Schredl M. Testing the involvement of the prefrontal cortex in lucid dreaming: A tDCS study. *Conscious Cogn* 2013;**22**:1214–1222.

Taube W, Mouthon M, Leukel C et al. Brain activity during observation and motor imagery of different balance tasks: An fMRI study. *Cortex* 2015;**64**:102–114.

Teng C, Kravitz DJ. Visual working memory directly alters perception. *Nat Hum Behav* 2019;**3**:827–836.

Voss U, Holzmann R, Hobson A et al. Induction of self-awareness in dreams through frontal low current stimulation of gamma activity. *Nat Neurosci* 2014;**17**:810–812.

Zabicki A, de Haas B, Zentgraf K et al. Imagined and executed actions in the human motor system: Testing neural similarity between execution and imagery of actions with a multivariate approach. *Cereb Cortex* 2017;**27**:4523–4536.

Zeman A, Dewar M, Della Sala S. Lives without imagery—Congenital aphantasia. *Cortex* 2015;**73**:378–380.

Zmigrod L, Garrison JR, Carr J et al. The neural mechanisms of hallucinations: A quantitative meta-analysis of neuroimaging studies. *Neurosci Biobehav Rev* 2016;**69**:113–123.